Gesture Recognition Using Depth Images

[Extended Abstract]

Bin Liang
Charles Sturt University
Boorooma Street
North Wagga NSW, Australia
bliang@csu.edu.au

ABSTRACT

This work presents an approach for recognizing 3D human gestures by using depth images. The proposed motion trail model (MTM) consists of both motion information and static posture information over the gesture sequence along the *xoy*-plane. By projecting depth images onto other two planes in 3D space, gestures can be represented with complementary information from additional planes. Accordingly 2D-MTM can be extended into 3D space in addition to the lateral scene parallel to the image plane to generate 3D-MTM. The Histogram of Oriented Gradient (HOG) is then extracted from the proposed 3D-MTM as the feature descriptor. The final recognition of gestures is performed through maximum correlation coefficient. The preliminary results demonstrate the average error rate decreases from 62.80% of baseline method to 21.74% after using the proposed approach on Chalearn gesture dataset.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*depth cues, motion, object recognition*

Keywords

Gesture recognition; depth images; 3D motion trail model; Kinect

1. INTRODUCTION AND MOTIVATION

Gesture is an important form for human interaction and communication. Therefore, many multi-modal interaction and computer vision applications have emerged thanks to a certain level of maturity achieved by sub-fields of machine intelligence, such as automatic environment surveillance, assisted living, video indexing and sport video analysis. As a major fulfillment of machine intelligence, gesture recognition has remained a prominent domain of research since the last three decades. In recent years, vision sensors such as video cameras are widely used for gesture recognition and research on vision-based interaction has been actively studied, because a vision sensor can provide an advanced interface and flexibility without extensive calibration and restriction of natural movement while wearable sensors are obtrusive and expensive [1]. The task of gesture recognition has made significant advances using ordinary vision sensor. Unfortunately, the successes have been limited to the use of RGB images captured by video cameras, ignoring the important information of depth. Depth information has long been regarded as an essential part of successful gesture recognition [2]. The Kinect camera, a cheap but quality depth sensor, provides depth information through collecting a sequence of depth images for human gesture. Depth images as 8-bit gray level for each pixel presents the distance between the captured object and camera. Therefore the motion ambiguity of the color camera, such as the huge color and texture variability induced by clothing, hair, skin and background, could be bypassed. This work studies the recognition of 3D human gestures from sequences of depth images.

2. RELATED WORK

Gesture recognition approaches can be categorized into three main categories: template based approaches, volumetric approaches and machine learning based approaches. Template based approaches [3, 4, 5] usually convert a gesture sequence into a static shape pattern, and then the extracted features are used to compare to the pre-stored prototypes during recognition. Template matching approaches are easy to implement and require less computational load. Volumetric approaches [6, 7] consider the whole gesture sequence as a 3D volume of pixel intensities instead of extracting features on a frame basis. These approaches have disadvantages of high computational cost and complexity. Machine learning techniques are employed for gesture recognition in recent years, such as SVM based methods [8], Bag-of-features methods [9] and HMM based methods [10], but they require offline learning phase and interactive computation. There have been many surveys on human gesture recognition and analysis [11, 12, 13]. Because of the less computational load and easier model representation, most of them have cited the motion history image (MHI) method [3] as one of the most popular methods. In the MHI, the silhouette sequence is condensed into gray scale image, while dominant motion information is preserved. Therefore, it can represent a gesture sequence in a compact manner. Besides, the MHI is not so sensitive to silhouette noises, like holes, shadows, and missing parts [14]. These advantages make MHI a suitable

ICMI'13, December 9–13, 2013, Sydney, Australia
ACM 978-1-4503-2129-7/13/12.
http://dx.doi.org/10.1145/2522848.2532188.

candidate for motion and gait analysis. The MHI expresses the motion history by the intensity of every pixel in temporal manner. However, the traditional MHI method has the limitation of the scalability because only lateral motion of the gesture is analyzed. Human gestures are performed in 3D space, which means MHI performed in 2D space may miss some motion information of the gesture performed in the real world.

In this work, a novel three dimensional motion trail model (3D-MTM) is proposed alternatively to overcome some of the drawbacks from the previous work. The 3D-MTM generated from the depth images explicitly models the temporal dynamics and static postures of the gestures in 3D space. For better modeling human gestures, our idea is to obtain the complementary motion information from other views in 3D space to compensate for lost gesture information in 2D space. 2D-MTM consists of both motion information and static posture information over the gesture sequence along the xoy-plane. After projecting depth images onto other two planes in 3D space, the 2D-MTM is then extensively combined with complementary motion information from additional two planes to generate a 3D-MTM that consists of six templates. Furthermore, feature vectors of the gestures are extracted from the proposed 3D-MTM for gesture recognition. The proposed method provides discriminantly informative representation for a gesture in 3D space containing disparity gesture information. The initial experimental results demonstrate the accuracy and effectiveness of the proposed 3D-MTM for human gesture recognition on Chalearn gesture dataset [15].

3. RESEARCH PLAN

My research project is titled gesture recognition using depth images. The research plan has been developed and included the following five tasks.

Task 1. Literature Review: The aim of this task is to investigate the current approaches in the area of gesture recognition and Human Computer Interaction (HCI), and then to give a critical reflective review on the collected reference papers.

Task 2. Research Proposal: According to the literature review, most of the approaches use 2D information obtained from ordinary cameras to perform gesture recognition. However, 2D gesture recognition methods eliminate the movement information along the z-axis. Thus, a 3D gesture representation model is proposed to recover the gesture information in 3D space for better gesture recognition.

Task 3. Methods Implementation and Analysis: Based on research proposal, the 3D gesture recognition method is implemented and will be improved in the following work. Furthermore, more details about research will be discussed and analyzed, more efficient features will be extracted and machine learning methods such as support vector machine will be used for feature classification.

Task 4. Dataset Analysis and Experiments: We plan to perform our experiments on the datasets provided by Chalearn Gesture Challenge. By now, they have released two datasets: one is for one-shot learning gesture recognition, and the other one is for multi-modal gesture recognition. After success on these datasets, we are going to build up our own dataset by using Kinect sensor.

Task 5. Paper Publishing and Thesis Writing: During the PhD study and research, outcomes about our research and contributions will be published. Finally, the thesis will be drafted and submitted.

4. PROPOSED APPROACH

The pipeline of proposed gesture recognition system has three layers as shown in Figure 1:

Preprocessing: Preprocessing consists of two steps: segmentation and smoothing. Segmentation is performed to segment human gesture regions from background. After that, smoothing is used in order to make sure there is less noise in depth images so that gesture analysis could not be negatively influenced by the noise.

3D Gesture Representation: In 3D gesture recognition, a more complete representation of the human body is required in order to characterize the movement properly. In this work, we propose 3D-MTM to represent human gestures in 3D space using depth images.

Feature Extraction and Classification: From the proposed 3D-MTM, feature vectors are computed for recognition purpose. Once the feature vectors for each gesture class are extracted, unknown gestures could be recognized through classification.

4.1 Preprocessing

With aim to segment the motion regions from the gesture sequences, background removal is an essential step for gesture recognition. We exploit depth information to segment motion regions from background. In depth images, the values of pixels belonging to background have a great difference from those belonging to the object. Utilizing the property, the motion region in a sequence can be easily segmented from the background by using Otsu's method [16] to classify the pixels. Besides, the depth images have other drawbacks, one of which is the noise at the edge of objects. With missing bits and a pretty serious flickering issue, noise in depth images resembles a type of salt and pepper noise. Motion information is sensitive to silhouette noise, smoothing of depth images is necessary. We adopt a 5×5 median filter [17] for spatial filtering to replace the pixel value with the median value of the sub-image. It helps remove the random noise.

4.2 2D Motion Trail Model

In the MHI, the gesture sequence is condensed into a single gray scale image, preserving dominant motion history information. It keeps a record of temporal changes at each pixel location, which decays over time [18]. However, in the presence of occlusions of body, or improper implementation of the update function, the MHI fails to cover most of the motion regions. In addition, the information of static posture history regions, repetitive movements and repetitive static postures are ignored in the MHI template. To improve the method, the proposed 2D-MTM in our work employs four templates, *i.e.* motion history image (MHI), average motion image (AMI), static posture history image (SHI) and average static posture image (ASI), to encode supplementary essential information of gestures to increase the robustness for representation.

Figure 2 illustrates the four templates of the proposed method performed on one gesture sequence. The images show that the MHI emphasizes recent motion, while the SHI emphasizes recent static posture information. Furthermore, AMI and ASI encode supplementary information of average motion and average static posture which both MHI and SHI

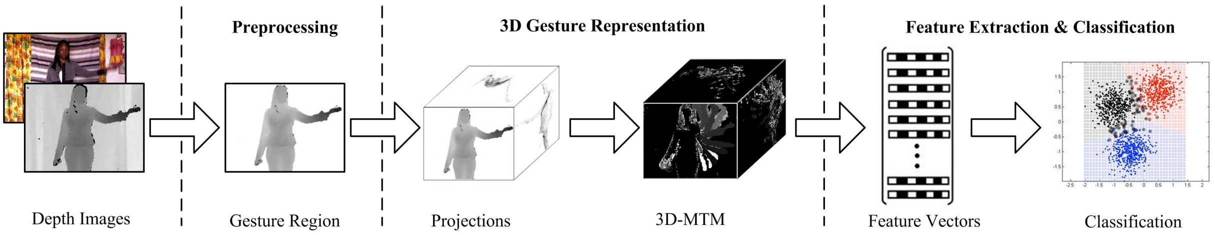

Figure 1: The pipeline of gesture recognition system

are poor to represent. Hence the combination of the four templates is complementary.

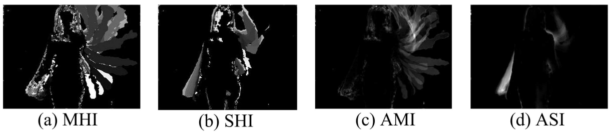

Figure 2: 2D motion trail model

4.3 3D Motion Trail Model

Our previous research on 2D-MTM based gesture representation has shown that the proposed model carries more essential gesture information in 2D space than traditional MHI method. However, the proposed 2D-MTM has some limitations. It can only encode the information induced by the lateral movement of the scene motion parallel to the image plane. As human bodies and motions are performed in 3D space, the information loss in the depth channel could cause significant degradation of the representation and discriminating capability for human gestures. With depth images, we can now extend the proposed 2D-MTM into 3D space, generating a 3D-MTM which is capable of encoding the motion information along other two additional planes (yoz-plane and xoz-plane) besides xoy-plane. Thus 3D-MTM uses disparity information of the gesture from xoy-plane, yoz-plane and xoz-plane, which can robustly discriminate each gesture using information from additional viewpoints with only one model. Figure 3 shows the 3D-MTM projections of one sample frame from a gesture sequence.

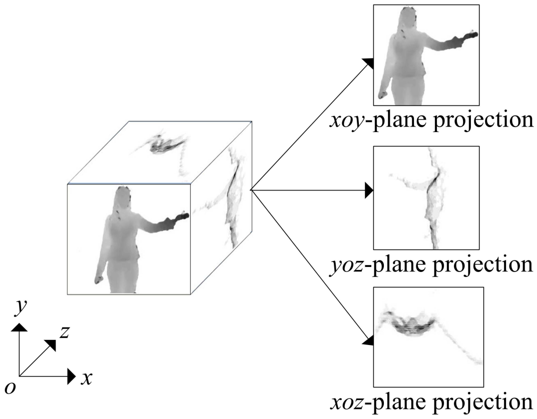

Figure 3: 3D-MTM projections of one sample frame.

The information from xoy-plane is dominant for the gesture and the projections onto yoz-plane and xoz-plane can be very coarse due to the resolution of the depth images, so only MHI templates are generated from the projections on yoz-plane and xoz-plane respectively. Thus, a gesture sequence from depth images can be represented using proposed 3D-MTM that consists of six templates: MHI, SHI, AMI, and ASI from xoy-plane, two additional MHI templates from yoz-plane and xoz-plane. The templates in 3D-MTM provide more supplementary information for the gesture than that of 2D space.

4.4 Feature Extraction and Classification

From the proposed 3D-MTM, feature vectors are computed for recognition purpose. In our preliminary work, HOG is employed to extract one feature vector from each template. Overall, six feature vectors are concatenated together for the matching process. Furthermore, we are planning to combine appearance and motion to obtain a more effective description of a wide variety of gestures. Through the proposed 3D gesture representation model, multidimensional histograms of the 3D flow (3D-HOFs) could be employed to catch motion information in the 3D space.

The initial experiment is performed on Chalearn gesture dataset that is designed for one-shot learning gesture recognition (each gesture class has only one training sample). Thus, maximum correlation coefficient as a nonparametric method is adopted by avoiding overfitting problem. In the following work, our plan is to perform the proposed model on the new dataset released by Chalearn Gesture Challenge. The new dataset has training examples for learning procedure. Therefore, some other learning based methods can be employed in the classification stage. Logistic regression, neural networks and SVMs will be employed in order to classify different gesture classes for performance comparison.

5. INITIAL EXPERIMENTS

The data in Chalearn gesture dataset [15] is recorded using a Kinect camera including both hand and arm gestures. It was used for a one-shot learning challenge of gesture recognition. The key aspect of the dataset is that each gesture class has only one training sample. Our experiments are performed on the first 10 data batches of the dataset, each of which is made of 47 gesture sequences.

Table 1 compares the average recognition error rate of the proposed 3D-MTM with results from other methods: baseline method [15], dynamic time warping (DTW), principle motion method [19], MHI method and proposed 2D-MTM. It can be observed that the proposed 3D-MTM shows a better performance which is competitive to the other methods. Our approach achieves 21.74% average error rate, which il-

Table 1: Comparison of the 3D-MTM with other methods

Method	Average error rate(%)
Baseline	62.80
Dynamic Time Warping	43.05
Principle Motion	37.42
MHI	37.64
2D-MTM (**ours**)	24.39
3D-MTM (**ours**)	**21.74**

lustrates that the 3D-MTM can be effectively adopted for gesture recognition.

6. CONCLUSIONS AND FUTURE WORK

In this paper, 3D-MTM is proposed as an effective solution to gesture recognition using depth images. The main idea of our proposed method is to compensate for the lost essential gesture information in the traditional MHI template, *i.e.* static posture information, repetitive movement information and repetitive static posture information. 3D-MTM is then extensively generated by combining motion information from yoz-plane and xoz-plane using depth images. The 3D-MTM encodes the discriminative motion trail information of a gesture, which is demonstrated on the one-shot learning Chalearn gesture dataset [15]. In comparison to recent work on the same dataset, 3D-MTM performs better with the average recognition error rate of 21.74%.

In the future work we would like to further explore other properties in depth images, motion descriptor and machine learning techniques for more accurate gesture recognition, which could better assist our method to conquer the ineffectiveness in discriminating hand and finger movements.

7. ACKNOWLEDGMENTS

My most heartfelt gratitude and appreciation is extended to my supervisor, Dr. Lihong Zheng, for always kindly finding time to support, question, and inspire my research.

8. REFERENCES

[1] Ronald Poppe. Vision-based human motion analysis: An overview. *Computer vision and image understanding*, 108(1):4–18, 2007.

[2] Allison Janoch, Sergey Karayev, Yangqing Jia, Jonathan T Barron, Mario Fritz, Kate Saenko, and Trevor Darrell. A category-level 3d object dataset: Putting the kinect to work. In *Consumer Depth Cameras for Computer Vision*, pages 141–165. Springer, 2013.

[3] Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(3):257–267, 2001.

[4] Md Ahad, Atiqur Rahman, JK Tan, HS Kim, and S Ishikawa. Temporal motion recognition and segmentation approach. *International Journal of Imaging Systems and Technology*, 19(2):91–99, 2009.

[5] Hongying Meng, Nick Pears, and Chris Bailey. A human action recognition system for embedded computer vision application. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–6. IEEE, 2007.

[6] Ivan Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[7] Yui Man Lui, J Ross Beveridge, and Michael Kirby. Action classification on product manifolds. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 833–839. IEEE, 2010.

[8] Edoardo Ardizzone, Antonio Chella, and Roberto Pirrone. Pose classification using support vector machines. In *Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on*, volume 6, pages 317–322. IEEE, 2000.

[9] Jutta Willamowski, Damian Arregui, Gabriella Csurka, Christopher R Dance, and Lixin Fan. Categorizing nine visual classes using local appearance descriptors. *illumination*, 17:21, 2004.

[10] Matthew Brand, Nuria Oliver, and Alex Pentland. Coupled hidden markov models for complex action recognition. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 994–999. IEEE, 1997.

[11] Jake K Aggarwal and Sangho Park. Human motion: Modeling and recognition of actions and interactions. In *3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004. Proceedings. 2nd International Symposium on*, pages 640–647. IEEE, 2004.

[12] M Ahad, JK Tan, HS Kim, and S Ishikawa. Human activity recognition: various paradigms. In *Control, Automation and Systems, 2008. ICCAS 2008. International Conference on*, pages 1896–1901. IEEE, 2008.

[13] Thomas B Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, 104(2):90–126, 2006.

[14] Md Atiqur Rahman Ahad, Joo Kooi Tan, Hyoungseop Kim, and Seiji Ishikawa. Motion history image: its variants and applications. *Machine Vision and Applications*, 23(2):255–281, 2012.

[15] ChaLearn. Chalearn gesture dataset. cgd2011, chalearn, california, 2011.

[16] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.

[17] Sangheon Park, Sunjin Yu, Joongrock Kim, Sungjin Kim, and Sangyoun Lee. 3d hand tracking using kalman filter in depth space. *EURASIP Journal on Advances in Signal Processing*, 2012(1):1–18, 2012.

[18] Tao Xiang and Shaogang Gong. Beyond tracking: Modelling activity and understanding behaviour. *International Journal of Computer Vision*, 67(1):21–51, 2006.

[19] Hugo Jair Escalante and Isabelle Guyon. Principal motion: Pca-based reconstruction of motion histograms. Technical report, Technical report, ChaLearn Technical Memorandum, June 2012. http://www. causality. inf. ethz. ch/Gesture/principal_motion. pdf, 2012.